THE 1996 BBN BYBLOS HUB-4 TRANSCRIPTION SYSTEM

Francis Kubala, Hubert Jin, †Spyros Matsoukas,
Long Nguyen, Rich Schwartz, John Makhoul

BBN Systems and Technologies, Cambridge MA 02138
†Northeastern University, Boston MA 02115

ABSTRACT

In this paper, we describe the BBN Byblos system used for the 1996 Hub-4 Partitioned Evaluation (PE) and Unpartitioned Evaluation (UE) tests. For the PE, we chose to ignore the segment feature labels that were given to the system as side-information so that our approach would generalize trivially to the UE. Moreover, we chose not to model specific channel conditions in the training because the observed gains were too small to warrant the additional system complexity required to support them. In the end, we estimated a single set of acoustic models from only 40 hours of broadcast news data. For the UE, the data was automatically segmented with a simple dual-gender phoneme recognizer that efficiently located pauses and changes in speakers' gender. After this preliminary stage of segmentation and gender-classification, our UE and PE systems were identical. We achieved a 30.2% word error rate on the PE test and 31.8% on the UE test - only a 5% relative degradation from our PE result.

1. INTRODUCTION

The 1996 Hub-4 tests evolved in a fashion that posed several new problems to researchers. The transcriptions accompanying the acoustic training data went through several iterations of refinement that didn't settle down until 2 weeks before the evaluation test period began. This required that we fix our development paradigm well before the training data had stabilized and then rapidly cut over to the new data when it arrived in its final form.

In addition, the total amount of Broadcast News (BN) data made available for acoustic training was only about 38 hours of speech. From past experience on large vocabulary tasks, this amount of within-domain acoustic data seemed rather small, especially since it was divided among a wide variety of conditions. Only about half of the BN data came from native speakers in a quiet environment over a wide-band channel. In response, we felt compelled to compare adaptation from a WSJ seed model estimated from 72 hours of clean speech to training directly on the 38 hours of mixed-condition BN data.

Finally, the PE and UE tests differed in the total side-information made available to the system, beyond the simple difference in segment boundaries. In the PE, the data was partitioned into segments having constant speaker, channel, and background conditions and the segments were given to the system with feature labels denoting these conditions. This test was conceived as a pedagogically useful breakdown of the real problem that would permit developers to concentrate on fundamental recognition problems without being diverted into dealing with pre-segmentation issues. But for systems intended to be evaluated on both the PE and UE tests, this test design required a choice between deploying the same system for both tests, or using the additional side-information for the PE and deploying a different system for the UE.

After conducting exploratory experiments to probe several different training paradigms, we opted for simplicity by training from scratch using only the BN data and by creating only a single set of condition-independent prior models. Furthermore, we chose to ignore the segment condition labels given in the PE test in order to focus on approaches that would be viable for the general UE problem. As a consequence, our PE system was identical to our UE system, save for an additional pre-segmentation step needed for the UE. This strategy permitted us to easily achieve a very small difference in performance between the two tests and demonstrated that tackling the whole transcription problem head on is possible now without diverting research away from the fundamental problems.

2. SYSTEM ARCHITECTURE

The 1996 Byblos Transcription System is organized into the following logical stages:

1. Segment and classify gender
2. Cluster the segments
3. Decode with Speaker-Independent (SI) models, to get transcriptions for adaptation
4. Adapt models to each cluster
5. Decode with Speaker-Adapted (SA) models, to produce the final answer

We create a set of general acoustic models that are specific to a particular stage. The models in the set are distinguished by their tying across HMM states, the training paradigm used, the gender of the training speakers, and the presence or absence of cross-word triphones.

In our Phonetically Tied Mixture (PTM) model, we allocate one mixture per phoneme. Each mixture has 256 components, resulting in about 12K Gaussians total. In our State-Clustered Tied Mixture (SCTM) model, 1000 mixtures are estimated with 64 components each. We create two SCTM models - one with only within-word triphones, and

one with cross-word triphones included. A pooled speaker training paradigm is used to create models for the SI stage. A Speaker-Adapted Training (SAT) paradigm [2] is used for the SA stage. Finally, separate models are estimated for each gender.

3. RECENT IMPROVEMENTS

3.1. 2-Pass N-best Decoder.

We have recently simplified our decoder strategy. A new 2-pass N-best decoder [7] has been implemented that is smaller and faster than our previous 4-pass decoder. This decoder uses a fast-match algorithm in the forward pass with PTM models, within-word triphones, and a bigram LM, producing a lattice of word ending times and scores. The backward pass uses SCTM models, within-word triphones, and a trigram LM, to produce the N-best hypotheses. After decoding, the N-best list is rescored with a cross-word SCTM model to produce the 1-best answer.

3.2. SNR-Dependent Normalization

We extended our simple Cepstral Mean Subtraction (CMS) normalization to accomodate a separate adjustment for speech and noise, in the manner of [1]. For each frame of data, we compute the probability of its being speech or noise based on frame energy. We estimate separate normalization vectors for speech and noise and shift each frame with a weighted combination of the two. In comparison to our old method, we observed a small gain on clean speech, and a small loss for noisy data.

3.3. Speaker-Adapted Training

The goal of SAT is to remove the variability among training speakers to achieve more compact HMM distributions [2]. Until recently, we had never tried to adapt more than 300 training speakers due to the prohibitive I/O requirements of our initial implementation. The BN training corpus contained over 2000 speakers, so we re-implemented SAT to handle large populations of speakers [6]. For the 1996 Hub-4 test, we adapted to 1000 speakers that had more than 20 seconds of speech. The remaining speakers (about 1400), accounting for about 15% of the training, were simply added as SI training. Due to lack of time, SAT models were only used in the final rescoring stage. Normally, we would use adapted SAT models for both passes of the final decode as well.

3.4. Acoustic Segmentation (UE)

For the UE test, the large monolithic input waveforms need to be cut at gender-change boundaries and classified as male or female, since our acoustic models are gender-dependent (GD). We also need to break the long segments into shorter ones for computational efficiency in the N-best stage. We accomplished both tasks with a dual-gender, PTM, context-independent phoneme decoder. Male and female HMMs were decoded in parallel in a single pass over the data, resulting in a sequence of time-stamped pauses and gender-tagged phones. The desired segments could then be produced by cutting the input at pause locations and gender changes indicated in the phone transcription. Boundary decisions were guided by several heuristics. No segment was permitted to be shorter than 2 seconds. And boundaries were not located within pauses shorter than 150 milliseconds, unless the hypothesized segment grew beyond about 10 seconds.

This simple model proved to work very well. It effectively rejected segments of pure music or noise by labeling them as pauses (non-speech intervals were included in the pause training). In fact, it was better at noise rejection than our PE approach. The gender classification was surprisingly stable at the phoneme level - the labels within a segment had a high degree of purity. Even so, it should be easy to improve on this simple model. Context-dependent phone models should always be better and we could measure the acoustic dissimilarity of segments adjacent to any pause to determine if a cut should be made at that point. By splitting dissimilar segments apart, we should improve our ability to cluster similar segments together for unsupervised adaptation.

3.5. Linguistic Segmentation

It seems reasonable to assume that the segmentation stage should produce segments containing word sequences that are consistent with the LM training. All of our LM training data is tagged with sentence boundaries, so we'd like to be able to chop segments with knowledge of where the likely sentence boundaries occur. We revived our work in linguistically-guided segmentation that we introduced in [5]. The basic idea is to change the LM so that the decoder can score a transition to a sentence end after every word hypothesis in the decoder. Then a decision to break the segment is made as a function of the sentinel score and the length of the pause at that time. But as we found previously, we have not been able to improve on the simple expedient of cutting at the longer pauses located by the standard decoder.

3.6. Speaker Clustering

The goal of speaker clustering is to group segments from the same speaker and condition together to improve the effectiveness of unsupervised adaptation. We have developed a fully automatic blind clustering algorithm [4] to accomplish this. We cluster segments (within each episode and gender) using a segment distance measure borrowed from our work in Speaker Identification [3]. A penalty is applied against the number of clusters created to establish a termination criterion in conjunction with the likelihood of the model at each stage of splitting. A positive bias is applied for segments that occur close in time, since they are more likely to be from the same speaker or condition. This approach worked as well as the ideal case of adapting with prior knowledge of speaker identity and signal condition.

4. HUB-4 ISSUES

4.1. Acoustic Training.

We used approximately 40 hours of speech for acoustic training. The 1996 Hub-4 training set contained data from 87 episodes, and we added the 10 episodes of 1995 Marketplace (MP) data to the training. We made no use of the F-condition labels in training; instead, all training data (within a gender) was pooled regardless of condition. This was done after we

observed that the gains for condition-specific models was too small to justify the additional system complexity required [8]. We did use the gender information in training, and for SAT training, we also used the speaker identity.

4.2. Transcription Validation.

Since the training data arrived so late, we needed to validate the transcriptions quickly and begin training immediately. We practiced on the preliminary release that became available on July 31, 1996. This release contained about 23 hours of actual speech. We aggressively rejected all troublesome segments reducing the training by the following percentages:

- any data labeled as *low fidelity* or high *music/noise* - 15%
- segments with OOV - 10%
- forced alignment failures - 1%
- failures in training - 2%

We ended up with 16 hours of usable speech for development. In the final release, which became available 2 weeks before the evaluation began, we again rejected 3% for those segments that failed alignment or training. But we included all low fidelity and high music or high noise data, and added all words to the lexicon before training.

4.3. Lexicon Design.

We created a word frequency list from Broadcast News (BN) and WSJ texts from the 1992-1996 period only. In coverage experiments on the Hub-4 development test, we observed no gain for using more data, and we saw no effect for weighting the data as a function of recency. We included all words found in the 1996 BN and 1995 MP acoustic training transcriptions, and then added words from the frequency list until we reached 45K words. At that point, coverage on the dev test was 99.1%. All new phonetic spellings were added by hand.

4.4. Language Model Training.

We used a total of 430M words for language model training from the following 5 LDC corpora:

- 131M 1992-96 BN, official Hub-4 release
- 254M 1988-94 WSJ
- 45M 1994-95 North American News
- 346K 1996 BN acoustic training
- 50K 1995 MP acoustic training

The final LM had 6M bigrams and 11M trigrams.

4.5. PE / UE System Differences.

There were only two differences between the systems we used for the PE and UE tests - the method used to segment and classify gender on the front end, and a filter on abnormally long words used on the back end. In all other respects, the two systems were identical. At the front end of the PE, we first decoded each given segment with male and female models. We classified the segments based on their recognition score, and then chopped the long segments at pauses located by the decoder. For the UE, we segmented at gender-changes, classified the segments for gender, and chopped them all in one step. At the back end of the UE only, we removed any word that was more than 2 times longer than had been observed in training. These giant words occur during periods of pure music or noise.

Our PE and UE systems could easily be completely identical. The phonetic gender segmenter is very accurate and much faster. And the long word filter works just as well on the PE. The only reason these systems differ at all is the lack of time we had for develoment.

5. EXPERIMENTS

As soon as we received and validated the 16 hours of BN training data, we estimated a PTM HMM from the data and compared it to our 1995 Hub-4 model. Last year's model was trained on 72 hours of WSJ data (close-talking microphone channel) and then adapted to the 10 episodes (4 hrs) of Marketplace training. In table 1 we show the WER for both systems on 4 episodes from the 1996 Hub-4 development test. For this experiment, both systems used SI PTM HMMs and within-word triphones. The new model is better

System	1995	1996
NPR The World	57.2	54.0
C-SPAN Washington Journal	56.2	48.7
NPR Morning Edition	42.0	39.8
Marketplace	31.3	28.8

Table 1: Comparison of two training paradigms: 1995 adapted WSJ and 1996 BN pooled-condition training

for each episode, despite its having only about one fifth as much training. Surprisingly, the new model is better for the Marketplace episode as well, even though last year's model was adapted to that very show. It's quite possible that the distant-mike channel of the WSJ corpus would perform better as a seed model for adaptation to BN. But at this point we decided to abandon the adaptive training paradigm that we used in 1995. And when the final release of the BN data arrived, increasing the BN training to 40 hours, the performance improved with a 6% relative reduction in WER.

We conducted additional experiments to determine how best to use the BN training data, given its highly variable composition. Is it better to create models for each of the conditions with supervised adaptation to the partitioned training data, or simply pool all the data BN data together and rely on unsupervised adaptation to the test to handle the variable conditions? These and related issues are discussed in [8].

In table 2, we show our PE test results for the November 1996 Hub-4 evaluation, broken out by condition. Recall that the PE test provides the system with segment boundaries and condition labels for each segment. We made no use of

Condition	SI	SAT adapted	relative gain
F0. prepared	23.8	21.6	9.2
F1. spontaneous	32.6	29.5	9.5
F2. low fidelity	38.0	32.7	14.0
F3. music	26.1	23.3	10.7
F4. noise	40.8	38.4	5.9
F5. non-native	36.1	31.8	11.9
FX. mixed	55.2	49.9	9.6
OVERALL	33.4	30.2	9.6

Table 2: PE evaluation result, for SI and SAT adapted recognition, by F-condition.

these segment labels, however. The results show gains in each condition for unsupervised adaptation to the test. The F2-condition, which includes the telephone data, enjoyed the largest gain. Still, the 9.6% overall gain for adaptation is rather small compared to improvements observed on other, less variable data such as WSJ.

Our 1996 UE evaluation result was 31.8% WER. We lost only about 5% relative for automatic segmentation. Last year, we determined that our segmentation algorithm was degrading performance by 21%. We tested our complete 1996 UE system, without any changes, on the 1995 Hub-4 evaluation test to measure our improvement for the year. The result was 26.5% WER compared to 42.7% last year. But WER alone doesn't tell the whole story. In contrast to the typical approaches adopted for the 1995 Hub-4 test, our new system has no speaker-dependent models, no channel-dependent models, and it is not show-dependent. Our new system is much better (in terms of WER), simpler, and more general than our previous one.

6. COMPUTATIONAL RESOURCES

We did the bulk of our processing on Silicon Graphics Indys with R4400 CPUs and 160 MB RAM. This is a 3 year old CPU, rated at about 90 SPEC92 for both integer and floating point performance. The SI decode (forward, backward, and rescoring) took about 80 times real-time. The SA decode ran faster at about 65 times real-time. The time required to adapt the models was significantly less than the decoding time and was dominated by I/O in our current implemetation. Computational requirements for training are given in [6].

7. CONCLUSIONS

We have demonstrated that the segmentation problem in broadcast news transcription is not a difficult one. Using a simple phonetic gender segmenter on the monolithic input waveform, we suffered only a 5% relative loss in performance from the PE test. Simple methods seem to work well enough. Therefore, the segmentation problem should not stand in the way of working on the whole problem, represented by the UE test.

In both training and decoding for the PE test, we made no use of the given F-condition labels. In training, data from all conditions was pooled. For test, we automatically identified speakers and channel conditions by a blind clustering procedure [4]. Also, though we found that we could achieve small improvements using condition-specific models, we considered the gain too small to justify the additional system complexity [8]. So the F-condition labels have no value for us in training, and in test, they are only useful for diagnostic purposes.

Finally, the 1996 Hub-4 results show that degradation due to channel, background, and speaking style conditions is secondary to the fundamental speech recognition error rate. Performance on clean, wideband, prepared speech from broadcast news is about 20%, which is completely unacceptable. That is the problem on which we intend to focus our efforts.

Acknowledgements

This work was supported by the Advanced Research Projects Agency and monitored by Ft. Huachuca under contract No. DABT63-94-C-0063. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the Government and no official endorsement should be inferred.

References

1. Acero, A., X. Huang, "Augmented Cepstral Normalization for Robust Speech Recognition", *Proceedings of IEEE Automatic Speech Recognition Workshop*, Snowbird UT, Dec. 1995, pp. 146-147.
2. Anastasakos, T., J. McDonough, R. Schwartz, "A Compact Model for Speaker-Adaptive Training", *Proceedings of ICSLP-96*, Philadelphia PA, Oct. 1996.
3. Gish, H., M. Siu, R. Rolicek, "Segregation of Speakers for Speech Recognition and Speaker Identification", *Proceedings of ICASSP-91*, Toronto, Canada, May 1991, vol. 1, pp. 701-704.
4. Jin, H., F. Kubala, R. Schwartz, "Automatic Speaker Clustering", *1997 DARPA Speech Recognition Workshop*, Chantilly VA, Feb. 1997, elsewhere this volume.
5. Kubala, F., T. Anastasakos, H. Jin, L. Nguyen, R. Schwartz, "Transcribing Radio News", *Proceedings of ICSLP-96*, Philadelphia PA, Oct. 1996.
6. Matsoukas, S., R. Schwartz, H. Jin, L. Nguyen, "Practical Implementations of Speaker-Adaptive Training", *1997 DARPA Speech Recognition Workshop*, Chantilly VA, Feb. 1997, elsewhere this volume.
7. Nguyen, L., R. Schwartz, "Efficient 2-Pass Nbest Decoder", *1997 DARPA Speech Recognition Workshop*, Chantilly VA, Feb. 1997, elsewhere this volume.
8. Schwartz, R., H. Jin, F. Kubala, S. Matsoukas, "Modeling the F-Conditions (or not)", *1997 DARPA Speech Recognition Workshop*, Chantilly VA, Feb. 1997, elsewhere this volume.